3Ю

УДК 004.934

О.А. Юхименко, В.В. Пилипенко, Р.А. Селюх
Міжнародний науково-навчальний центр інформаційних технологій та систем,
м. Київ, Україна
Україна, 03680, м. Київ, просп. Акад. Глушкова, 40

Адаптація до голосу нового диктора на прикладі спонтанного мовлення з корпусу АКУЕМ

O.A. Yukhymenko, V.V. Pylypenko, R.A. Selyukh
International Research/Training Centre for Information Technologies and Systems, Kyiv, Ukraine
Ukraine, 03680, Kyiv, prosp. Akad. Hlushkova, 40

Adaptation to New Announcer Voice for Spontaneous Speech from AKUEM Speech Corpus

А.А. Юхименко, В.В. Пилипенко, Р.А. Селюх
Международный научно-учебный центр информационных технологий и систем,
г. Киев, Украина
Украина, 03680, г. Киев, просп. Акад. Глушкова, 40

Адаптация к голосу нового диктора на примере спонтанной речи из корпуса АКУЕМ

Стаття присвячена питанням адаптації до голосу нового диктора попередньо створених систем пофонемного розпізнавання мовлення. Представлені результати трьох експериментів, проведених з використанням даних мовленнєвого корпусу АКУЕМ. Надається порівняльний аналіз з результатами попередніх досліджень з адаптації.
Ключові слова: моделі фонем, адаптація, розпізнавання, лінійні перетворення, класи регресії, навчання.

This article is devoted to the problems of adaptation to new announcer voice for speech recognition systems. The results of three adaptation experiments based on AKUEM speech corpus are described. Comparison with previous experiments are discussed.
Key words: models of phonemes, speaker adaptation, recognition, linear transformations, classes of regression, training.

Статья посвящена вопросам адаптации к голосу нового диктора предварительно созданных систем пофонемного распознавания речи. Представлены результаты трех экспериментов, проведенных с использованием данных речевого корпуса АКУЕМ. Приводится сравнительный анализ с результатами предшествующих исследований по адаптации.
Ключевые слова: модели фонем, адаптация, распознавание, линейные преобразования, классы регрессии, обучение.

Вступ

У попередніх роботах була проведена серія експериментальних досліджень з адаптації, застосовані різні підходи [1], [2]. Слід зазначити, що вони були проведені в рамках пофонемного послівного розпізнавання. Всі диктори, записи котрих використовували в експериментах, наговорювали визначені певні слова, які апроксимують

фонетичне розмаїття української мови. При цьому слова вимовлялися загалом розбірливо, в нормальному темпі, окремо одне від одного. Диктори базового кооперативу вимовили більш ніж дванадцять тисяч реалізацій слів у загальній навчальній вибірці. Розпізнавання було послівним. Використовувалося два достатньо якісних мікрофони, умови запису відповідали офісним. Словник використовувався невеликий – біля 2,5 тисячі слів. Кількість дикторів також була невеликою – 67. У даній роботі представлені результати експериментальних досліджень, котрі були отримані дещо в інших умовах і не з окремими словами, а зі злитим, здебільшого спонтанним мовленням.

Метою роботи є продовження досліджень з адаптації в більш складних умовах роботи з мовленнєвим матеріалом.

Лінійні перетворення при адаптації акустичних моделей

При створенні системи розпізнавання сигналів мовлення необхідно провести процедуру навчання розпізнаванню. При пофонемному розпізнаванні кожна фонема має свою акустичну генеративну модель, котра являє собою певну кількість станів з певними переходами між ними [1]. При цьому кожний стан моделі має свої ймовірнісні параметри – середній вектор спостереження $\mu = [\mu_1, \mu_2, ..., \mu_n]^T$ та коваріаційну матрицю Σ розмірністю $n \times n$, де n – розмірність вектора первинних ознак сигналу. Ці μ та Σ є параметрами n-вимірного нормального закону розподілу. Стан моделі може задаватися декількома параметрами (парами), то тоді говорять, що стан описується сумішшю гаусіанів (нормальних розподілів). Проведення процедури навчання передбачає конкретне обчислення за допомогою ітераційних процедур саме цих ймовірнісних параметрів для всіх фонем у системі розпізнавання. Для двох систем розпізнавання, навчених на двох різних дикторів, ці ймовірнісні параметри будуть різнитися між собою, чим і пояснюється незадовільна точність розпізнавання якогось диктора на чужій системі.

Але цілком можливо обчислити лінійні перетворення, які переводять початкові середні вектори та коваріаційні матриці опорного диктора або кооперативу дикторів у середні вектори та коваріаційні матриці нового диктора. Лінійне перетворення для середніх векторів записується у вигляді:

$$\hat{\mu} = W\xi , \tag{1}$$

де $\hat{\mu}$ – середній вектор нового диктора, W є матрицею, розмірністю $n \times (n + 1)$, ξ – середній розширений вектор опорного диктора,

$$\xi = [1, \mu_1, \mu_2, ..., \mu_n]^T . \tag{2}$$

Лінійне перетворення коваріаційних матриць записується у вигляді:

$$\hat{\Sigma} = H\Sigma H^T , \tag{3}$$

де H – матриця перетворення коваріаційної матриці Σ опорного диктора, розмірністю – $n \times n$.

Щоб покращити гнучкість процесу адаптації, можна визначити відповідну множину базових класів, яка залежатиме від кількості доступних адаптаційних даних [3]. Якщо доступна мала кількість адаптаційних даних, то тоді буде генеруватися загальне адаптаційне перетворення. Загальне перетворення застосовується до кожної компоненти гаусіанів у множині моделей. Одначе, якщо адаптаційних даних стає більше, то можливо покращити адаптацію шляхом збільшення кількості перетворень. Тоді

кожне перетворення стає більш конкретним й застосовується до певної групи гаусіанів. Наприклад, гаусіани можуть бути згруповані в широкі фонетичні класи: пауза, голосні, назальні, фрикативні тощо. В цьому випадку адаптаційні дані повинні використовуватися для побудови більш конкретних перетворень широких класів, щоб застосувати ці перетворення до цих угруповань.

Зв'язування кожного перетворення через множину компонентів суміші дозволяє адаптувати й ті розподіли, для котрих узагалі не було спостережень. У такому процесі всі моделі можуть бути адаптовані й адаптаційний процес динамічно покращується, як тільки з'являється більше адаптаційних даних.

Дерево класів регресії побудовано таким чином, щоб об'єднати компоненти, котрі близькі в акустичному просторі, й, таким чином, схожі компоненти будуть перетворюватися схожим способом. Зазначимо, що дерево побудовано з використанням індивідуальної дикторонезалежної множини моделей фонем, а значить – не залежить від будь-якого нового диктора. Термінальні вузли або листки дерева визначають кінцеві групи компонентів й називаються базовими класами (класами регресії). Кожний гаусіан у множині моделей фонем належить до одного певного базового класу.

На рис. 1 наведено простий приклад бінарного дерева регресії з чотирма базовими класами, позначеними як $\{C_4, C_5, C_6, C_7\}$. На діаграмі зображено неперервні стрілки та неперервні околи й це означає, що адаптаційних даних, пов'язаних із цим класом, достатньо для побудови матриць перетворення. Пунктирні стрілки та околи позначають класи, для яких недостатньо адаптаційних даних. У цьому прикладі вузли 6 та 7 не мають достатньо даних; але у вузлі 3, що є батьківським для 6 та 7, даних достатньо. Аналогічно для вузлів 5 та 2. Кількість даних, що визначається як достатня (поріг), встановлюється як опція вручну в програмі.

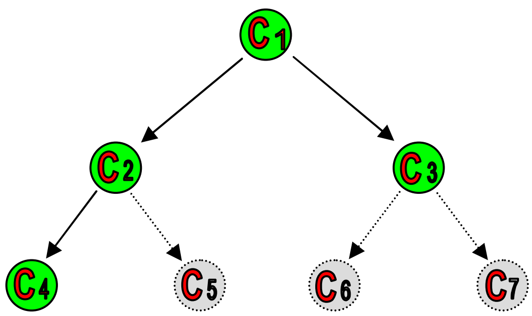


Рисунок 1 – Бінарне дерево регресії

Перетворення генеруються тільки для тих вузлів, котрі:

1) мають достатньо даних;

2) є або термінальними вузлами (тобто базовими класами), або мають нащадків з недостатньою кількістю даних.

У прикладі, котрий наводиться на рис. 1, перетворення генеруються лише для вузлів регресії під номерами 2, 3 та 4, й ці перетворення позначимо відповідно W_2, W_3 та W_4.

Звідси, коли потрібно мати перетворену множину моделей фонем, матриці перетворення (для середніх та дисперсій) застосовуються до компонентів гаусіанів у кожному базовому класі наступним чином:

$$\begin{Bmatrix} W_2 \to \{C_5\} \\ W_3 \to \{C_6, C_7\} \\ W_4 \to \{C_4\} \end{Bmatrix} .$$

Тут цікаво відзначити, що випадок загальної адаптації, схожий на випадок, коли дерево має лише один кореневий вузол.

Експериментальна база

Як було зазначено у вступі, в даній роботі експерименти проводилися переважно зі спонтанним мовленням. Воно полягає в тому, що диктори, записи котрих використовували в експериментах, говорили вільно або читали, не спеціально для експериментів, порядок слів у їхній мові був вільний, деякі слова вони повторювали й не завжди повністю, не завжди ясно й чітко, говорили з різним ступенем емоційності, в різному темпі, при цьому мовлення було злитим. Розпізнавання також проводилося для злитого мовлення. Каналів запису було багато, вони різнилися між собою за характеристиками. Записи дикторів були не однакового обсягу – від коротких за часом до довгих. Використовувалися записи з теле- та радіоефіру. Всі ці записи були зібрані в так званий корпус АКУЕМ – акустичний корпус українського ефірного мовлення [4]. В цьому корпусі словник налічував 71 545 словоформ, близько 60 годин аудіозаписів, у котрих міститься мовлення біля 2000 дикторів. Слід зазначити, що диктори говорили й такі слова, котрих не було в словнику взагалі, на відміну від [1]. Це ускладнювало ситуацію тим, що автоматично понижувало надійність розпізнавання. Більшість дикторів представлена короткими записами, тоді як у 150 дикторів довжина записів становить більш як 10 хвилин. З усього вищесказаного випливає, що, взагалі, умови для розпізнавання в даному випадку менш сприятливі, ніж у попередніх дослідженнях.

Кількість фонем, як і в попередніх дослідженнях, становила 55 елементів. Фонеми моделюються трьома станами Марківського ланцюгу без пропусків.

Попередні експериментальні дослідження для визначення значення порогу достатності адаптаційних даних

Взагалі, було проведено три експерименти з, відповідно, трьома різними Контрольними групами дикторів.

Контрольна група № 1 складалася з дикторів, котрі брали участь у навчанні. Тобто, записи промов цих дикторів були розділені на дві частини: записи з першої частини повністю використовувалися при навчанні системи розпізнавання (це була навчальна вибірка (НВ)), записи з другої частини використовувалися для тестування та адаптації (це була незалежна вибірка (НезВ) цих дикторів). Мета цього експерименту – експериментально з’ясувати, як залежать результати адаптації від кількості лінійних перетворень, котрі застосовуються при цій самій адаптації. Тобто, кількість адаптаційних даних не змінювалася, АВ залишалася тою самою, а змінювалось вручну

значення порогу достатності даних у дереві класів регресії. Чим більше це значення, тим менше буде лінійних перетворень на всю систему при адаптації. Приймалося 4 різних значення порогу – 2000, 1000, 500, 200. Будувалися різні дерева класів регресії – з 1, 2, 3, 4, 6, 8, 10, 13, 16, 20, 25 та 30 термінальними вузлами. Для кожного дерева, в залежності від значення порогу, обчислювалася різна кількість лінійних перетворень. Попутно необхідно було з'ясувати питання, в якому випадку результати адаптації будуть кращі: коли адаптаційну вибірку (АВ) брати з НВ, або коли з НезВ? Результати даного експерименту зображені на рис. 2.

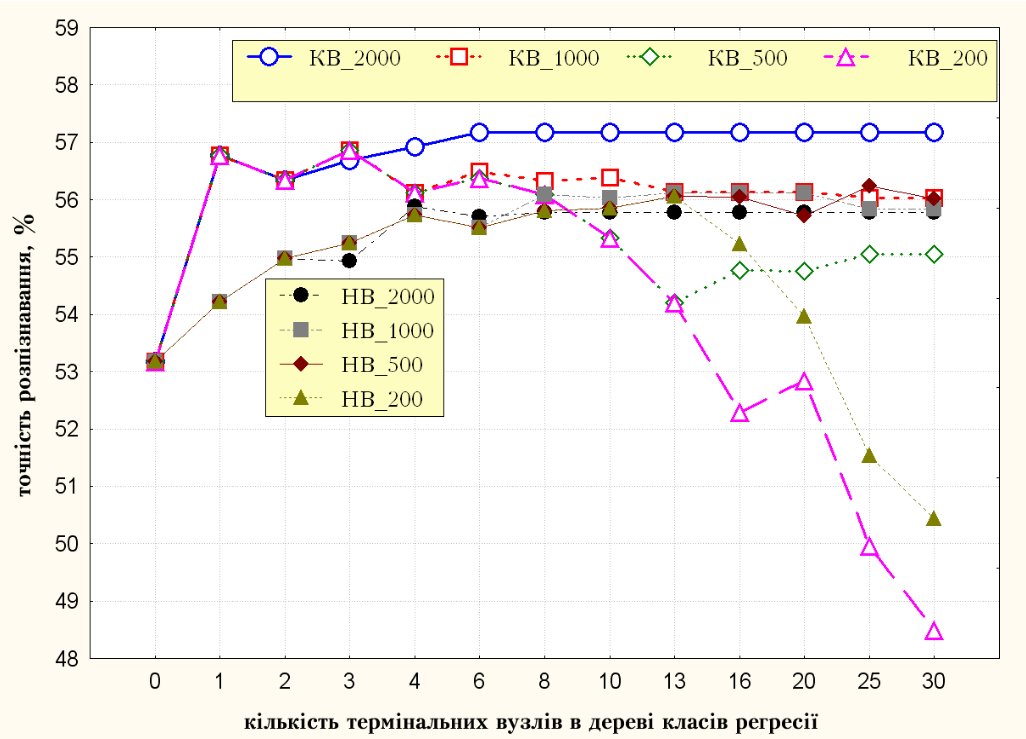


Рисунок 2 – Усереднена точність розпізнавання дикторів
із контрольної групи № 1 до та після адаптації

Пояснення: КВ_2000 – це значить, що АВ вибиралася з НезВ, значення порогу 2000; НВ_500 – АВ вибиралася з НВ, значення порогу 500. Коли кількість термінальних вузлів – 0, то це означає, що розпізнавання проводилося без адаптації. Досить ясно видно, що результати адаптації кращі, коли АВ вибирають з НезВ (при порогах 2000 та 1000), при порогах 200 та 500 отримуємо досить непевний результат. Виходило, що просте збільшення кількості перетворень (від пониження порогу) без збільшення обсягу АВ не призводить до автоматичного покращення розпізнавання. Можна констатувати, що збільшення точності розпізнавання при виборі АВ з НезВ сягає майже 4% (при порозі 2000), при виборі АВ з НВ сягає майже 3% (при порозі 500, 1000, 2000). Результати адаптації при виборі АВ з НВ менш розкидані (окрім порогу в 200). Дослідження проводилися при кількості гаусіанів у сумішах станів моделей фонем – 16.

У другому експерименті контрольна група № 2 складалася з дикторів, котрі не брали участі в навчанні. Тобто, записи промов цих дикторів не використовувалися при навчанні системи розпізнавання, вони мали лише НезВ. Мета – експериментально з'ясувати, чи будуть результати адаптації для групи, що не брала участі в навчанні, кра-

щими, ніж для групи, котра брала участь у навчанні. Одночасно необхідно було з'ясувати питання: як залежать результати адаптації при збільшенні кількості гаусіанів у сумішах станів моделей фонем? Оскільки в попередньому експерименті при значенні порогу 200 отримували незадовільний результат, то тут його не використовували. Дерева класів регресії – ті самі. Результати даного експерименту зображені на рис. 3.

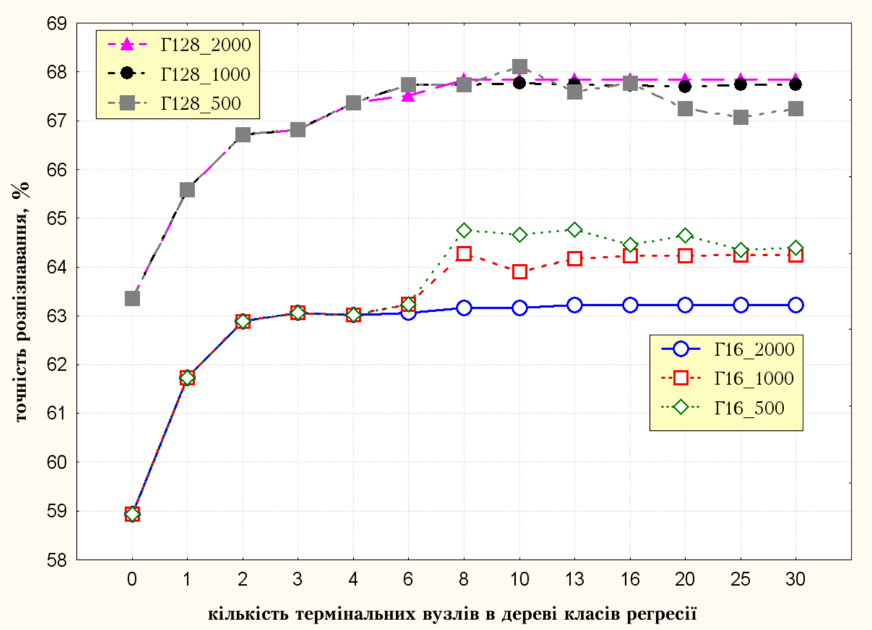


Рисунок 3 – Усереднена точність розпізнавання дикторів із контрольної групи № 2 до та після адаптації при 16 та 128 гаусіанах у моделях фонем

Пояснення: Г128_2000 – це значить, що гаусіанів в моделях фонем 128, значення порогу 2000. Чітко видно, що при 128 гаусіанах точність розпізнавання вища як до, так й після адаптації, результати менш розкидані.

Зростання точності – до 4,5% (поріг 2000) при 128 гаусіанах, до 5,5% (поріг 500, 1000) при 16 гаусіанах.

Порівнюючи з результатами адаптації першого експерименту можна зробити висновок, що при 16 гаусіанах результати адаптації покращилися – 5,5% проти 4%, відносне покращення також більше, хоча при цьому говорити про видатну різницю не доводиться.

Результати експериментальних досліджень на матеріалі виступів депутатів Верховної Ради України

Контрольна група № 3 складалася з дикторів, котрі також не брали участі в навчанні. Ці диктори – депутати Верховної Ради України (записи їхніх промов також знаходяться в АКУЕМ). Вони говорили зі специфікою парламентських промов і зі специфікою записів цих промов у парламентській залі. Мета – знову-таки експериментально з'ясувати, чи будуть результати адаптації для групи, що не брала участі в навчанні, кращими, ніж для групи, котра брала участь у навчанні. Також була поставлена задача: проводити адаптацію не для однієї певної АВ для кожного диктора, а для декількох різних за обсягом АВ, щоб оцінити якість адаптації в залежності від обсягів АВ та поставити

дикторів у рівні умови. АВ для всіх дикторів обиралися обсягом в 30, 60 та 90 секунд. Дерев класів регресії було побудовано трохи менше. Результати даного експерименту зображені на рис. 4, 5, 6.

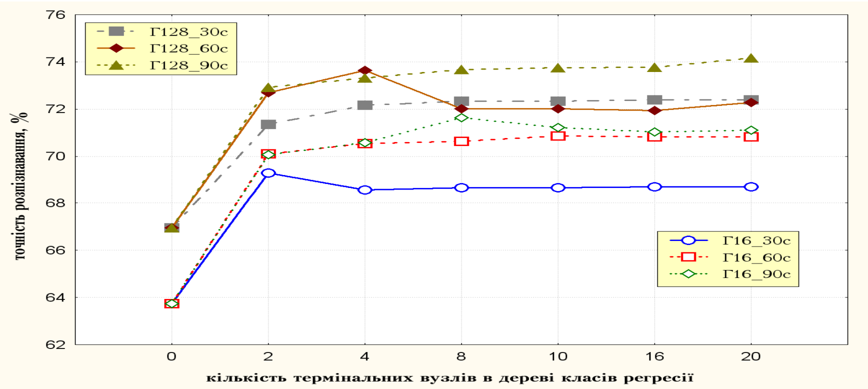


Рисунок 4 – Усереднена точність розпізнавання дикторів із контрольної групи № 3 до та після адаптації при 16 та 128 гаусіанах у моделях фонем, при значенні порога 500

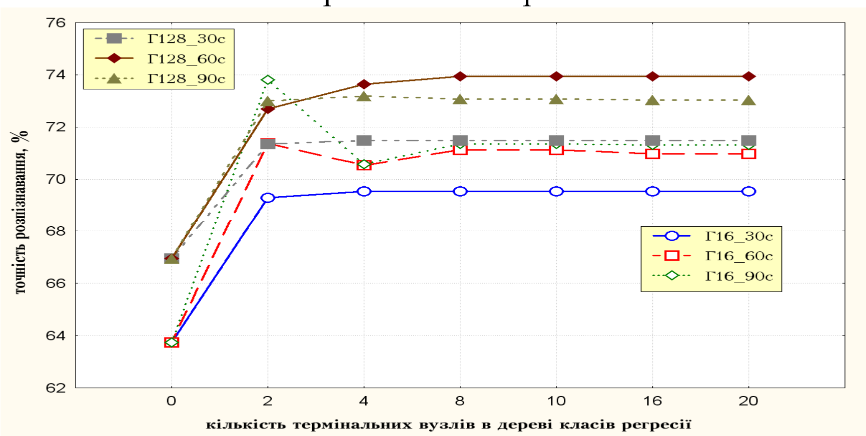


Рисунок 5 – Усереднена точність розпізнавання дикторів з контрольної групи № 3 до та після адаптації при 16 та 128 гаусіанах у моделях фонем при значенні порога 1000

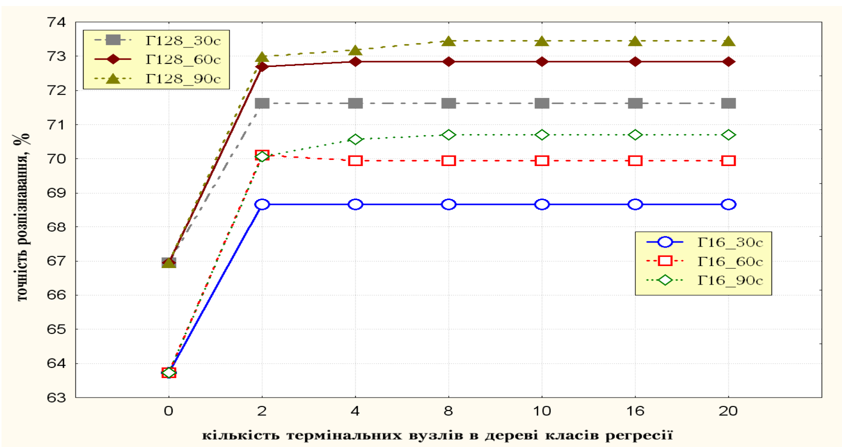


Рисунок 6 – Усереднена точність розпізнавання дикторів із контрольної групи № 3 до та після адаптації при 16 та 128 гаусіанах у моделях фонем при значенні порога 2000

Пояснення: Г16_60с – гаусіанів у моделях фонем 16, обсяг АВ – 60 секунд.

З рисунків видно, що при збільшенні обсягу АВ росте точність розпізнавання після адаптації. Результати при АВ в 30 секунд гірші за результати при АВ в 60 та 90 секунд, у свою чергу АВ в 60 та 90 секунд при 128 гаусіанах і порозі 500 та 1000 дають між собою зворотній результат. Для подальших експериментів було обрано значення порогу достатності даних 2000, оскільки майже в усіх випадках при ньому досягається найбільша точність і результати більш стабільні при зміні кількості класів у дереві регресії. В цьому випадку при 128 гаусіанах маємо зростання точності після адаптації від 4,5% (при 30с) до 6,5% (при 90с), при 16 гаусіанах – від 5% (при 30с) до 7% (при 90с). Спостерігається збільшення точності розпізнавання порівняно з контрольною групою № 1.

Висновки

Отже, експерименти наявно показали доцільність застосування адаптації до голосу нового диктора.

Було з'ясовано, що при збільшенні гаусіанів (тут конкретно від 16 до 128) спостерігається покращення точності розпізнавання. Одначе після адаптації більший ріст точності мав місце саме при 16 гаусіанах.

Для дикторів, що брали участь у навчанні, ріст точності розпізнавання після адаптації був дещо більший тоді, коли АВ вибиралася з НезВ. Для дикторів, що не брали участі в навчанні, ріст точності розпізнавання після адаптації був дещо більший у порівнянні з дикторами, що брали участь у навчанні.

Зменшення значення порогу призводить до збільшення кількості лінійних перетворень. Експерименти показали, що просте зменшення значення порогу для збільшення кількості перетворень взагалі не призводить до автоматичного покращення точності. Це стається, очевидно, з причини погіршення статистик внаслідок зменшення кількості спостережень при зменшенні значення порогу.

Експеримент № 3 показав, що, взагалі, бажано брати АВ обсягом не менш за 60 секунд, хоча й 30 секунд давали зростання точності. Збільшення АВ покращує результати адаптації, принаймні до якогось моменту. Задача на майбутнє – з'ясувати, коли наступає цей момент, тобто такі обсяги АВ, що подальше нарощування АВ не дає збільшення точності розпізнавання.

Експерименти представили, що ми маємо впевнене зростання надійності розпізнавання після адаптації біля 4 – 5%, хоча в певних варіантах (при АВ в 90с) було й більше. У роботі [1] початкове розпізнавання було помітно більшим – майже 90%, середня надійність розпізнавання самих дикторів базового кооперативу сягала 94,32%. Після адаптації тоді було досягнуто до 6% зростання надійності, отже відносне покращення було також суттєво більшим. Але все це відбулося, безсумнівно, внаслідок загалом більш сприятливих умов для розпізнавання.

Література

1. Сажок М. Адаптація акустичних моделей фонем до голосу диктора для пофонемного розпізнавання ізольованих слів української мови / М. Сажок, Р. Селюх , О. Юхименко // Штучний інтелект. – Донецьк, 2009. – № 4. – С. 230-233.
2. Сажок М. Адаптація до голосу диктора на основі гендернозалежних акустичних моделей фонем для української мови / М. Сажок, Р. Селюх, О. Юхименко. – Оброблення сигналів і зображень та

розпізнавання образів : Десята Всеукраїнська міжнародна конференція. – Київ, 2010. – С. 59-62.
3. Young S.J. HTK Book, version 3.1 / Young S.J. [et al]. – Cambridge University, 2002. – 355 p.
4. Створення акустичного корпусу українського ефірного мовлення / [Н.Б. Васильєва, В.В. Пилипенко, О.М. Радуцький та інш.]. – Оброблення сигналів і зображень та розпізнавання образів : Десята Всеукраїнська міжнародна конференція. – Київ, 2010. – С. 55-58.

Literatura

1. Sazhok M. Adaptatsija akustychnykh modelej fonem do holosu dyktora dlya pofonemnogo rozpiznavannya izolyovanykh sliv ukrajinskoji movy / M. Sazhok, R. Selyukh , O. Yukhymenko // Shtuchnyj intelekt. – Donetsk, 2009. – № 4. – s. 230-233.
2. Sazhok M. Adaptatsija do holosu dyktora na osnovi gendernozalezhnykh akustychnykh modelej fonem dlya ukrajinskoji movy / M. Sazhok, R. Selyukh , O. Yukhymenko. – Obroblennya sygnaliv i zobrazhen ta rozpiznavannya obraziv : Desyata Vseukrajinska mizhnarodna konferenciya. – Kyiv, 2010. – s. 59-62.
3. Young S.J. HTK Book, version 3.1 / Young S.J. [et al]. – Cambridge University, 2002. – 355 p.
4. Stvorennya akustychnoho korpusu ukrajinskoho movlennya / [N.B. Vasyl'eva, V.V. Pylypenko, O.M. Radutskyj et al]. – Desyata Vseukrajinska mizhnarodna konferenciya. – Kyiv, 2010. – s. 55-58.

RESUME

O.A. Yukhymenko, V.V. Pylypenko, R.A. Selyukh

Adaptation to New Announcer Voice for Spontaneous Speech from AKUEM Speech Corpus

The article is continuation of series of experiments on adaptation to voice of new announcer of the preliminary created systems of phoneme recognition. If in previous works as units of speech signals were the isolated words, in these experiments information was used from the vocal corpus of AKUEM (mostly, spontaneous speech).

The presented results of three experiments deal with the different sizes of adaptation sets and parameters of adaptation.

A comparative analysis is given with the results of previous adaptation researches.

The results of experiments show an improvement reliability of recognition after adaptation to voice of a new speaker.

Стаття надійшла до редакції 09.04.2013.